Using Few Clues can Compensate the Small Amount of Resources Available for Word Sense Disambiguation

Claude de Loupy* & Marc El-Bèze*

* Laboratoire Informatique d'Avignon B.P. 1228, Agroparc, 339 chemin des Meinajaries 84911 Avignon, cedex 9, FRANCE {marc.elbeze, claude.de.loupy}@lia.univ-avignon.fr

Abstract

Word Sense Disambiguation (WSD) is considered as one of the most difficult tasks in Natural Language Processing. Probabilistic methods have shown their efficiency in many NLP tasks, but they imply a training phase and very few resources are available for WSD. This paper aims at showing how to make the most of size-limited resources in order to partially overcome the knowledge acquisition bottleneck. Experiments are performed within the SENSEVAL test framework in order to evaluate the advantage of a lemmatized or stemmed context over an original context (inflected forms as they are observed in the rough text). Then, we measure the precision improvement (about 6 %) when looking at the inflected form of the word to be disambiguated. Lastly, we show that it is possible to reduce the ambiguity if the word to be disambiguated has a particular inflected form or occurs as part of a compound.

1. Introduction

Any system involving text processing has to cope with several difficulties inherent in natural language. Polysemy and synonymy are two of the most important problems faced by several applications (Ide & Véronis, 1998) like Document Retrieval (DR), Machine Translation and Text Classification. Studies have been carried out in order to point out the influence of polysemy on IR system performances (Gonzalo *et al.*, 1998). A Word Sense Disambiguation (WSD) component is generally used for this purpose and is expected to be as efficient as possible.

Corpus-based probabilistic methods have shown their efficiency in many Natural Language Processing (NLP) tasks like speech processing or Part Of Speech (POS) tagging. Several papers report the use of such methods for WSD (Yarowski, 1992; Segond *et al.*, 1997; Loupy *et al.*, 1998). However, even if the results reported are encouraging, it has been underlined that there are too few resources available for WSD. Their sizes are far smaller than the ones of those used for other NLP tasks. One challenge in WSD is to cope with the data sparseness problem.

WSD is a very difficult task and supervised methods require a lot of training data in order to assign a sense to a word in context with enough confidence. This is the well known phenomenon called *knowledge acquisition bottleneck* (Karov & Edelman, 1996).

This paper aims at showing how to make the most of size limited-resources. We intend to show that using few clues can help systems to partially overcome the lack of training. We first present the data and method used for the experiments. Several features can be used in order to perform a WSD task. In a first set of experiments (section 3), we compare the performances obtained when using a lemmatized or a stemmed context, or the original context that is inflected forms exactly as they appear in the text (with uppercase or lowercase). The second set shows that the use of the inflected form for the word to be tagged can bring a lot of information in order to determine its sense (section 4). The last set demonstrates that it is possible to dramatically reduce the ambiguity of a word taking into account its inflected form and the previous or following word (sections 5 and 6).

2. Data and method used

2.1. SENSEVAL

SENSEVAL (Kilgarriff, 1998; Kilgarriff & Palmer, 2000) was the first open evaluation campaign for WSD. 25 systems were tested: 18 for English, 5 for French and 3 for Italian . We participated for English and French (Loupy *et al.*, 2000).

Thanks to SENSEVAL, appropriate resources are now available in order to compare different approaches used for WSD¹. They are based on the HECTOR (Atkins, 1993) lexical database. 15 nouns, 13 verbs, 8 adjectives and 5 indeterminate words (the POS is not given) constitute the test framework. For some words, there is no training data.

The following experiments are applied on nouns and, since the method requires a training corpus, 12 nouns are considered (and not 15).

NOUN	TAGS	SAMPLES	TESTS
accident	8	1265	267
behaviour	3	1004	279
bet	15	146	274
excess	7	201	186
float	12	83	75
giant	7	367	118
knee	21	460	251
onion	4	41	214
promise	8	616	113
sack	11	113	82
scrap	14	67	156
shirt	8	556	184
Total	118	4919	2199

Table 1: Test data

In order to train the models, we have used the examples given by the dictionary and the training corpus. Table 1 shows the number of senses (TAGS), the total

¹ Resources are available at

http://www.itri.brighton.ac.uk/events/senseval

number of samples (SAMPLES) and the number of tests (TESTS) for each noun to be tagged.

2.2. Preparation of the data

The method uses a very short context of words. The window is only 2 words before and 2 words after the word to be tagged. Training and test corpora are grammatically tagged. This is necessary for lemmatization and useful in order to eliminate particular tokens. Words belonging to some grammatical classes are not taken into account in our WSD system: determiners, adverbs and adjectives.

It is possible to unify words (months for instance) so strongly related that replacing one by another does not modify the sense of the test. Hence, January, February, etc. are replaced by MONTH and Monday, etc. by DAY. Moreover, CD stands for any number, PRP for pronouns, NNPL for a location (like Paris), NNPG for a group (a company, etc.), NNPP for the name of a person and UNK for an out-of-vocabulary-word if its initial letter is an uppercase. These substitutions are intended to decrease the variability of the context in which a given word sense can be found.

2.3. The KNN method

A very simple system can be used in order to disambiguate a word in context (El-Bèze, *et al.*, 2000). The K Nearest Neighbors (KNN) method provides a simple way to assign a semantic tag. For a word occurring in the context of the test to be disambiguated, it is interesting to check whether this word also appears in the context of any example of the training corpus. Each training sample sharing a same word at the same place (found in the test context) votes for its semantic tag. An example has a number of votes equal to the number of words shared in the same position with the test.

When the poll is done, the test is tagged with the sense which has received the greater number of votes. If several tags get the best score, or if no sample matches the test, the most frequent one among possible senses is chosen. When there is equality in samples and frequency, all the tags having the same score are kept and each one is assigned the same weight (the sum of these weights is 1). Since a tag is always proposed by the system, recall is equal to precision.

3. Which linguistic level is useful for WSD?

What we are looking for is clues that help the system to disambiguate the sense of a word in context. It is well known that taking into consideration the POS really improves the performances (Wilks & Stevenson, 1996). In this paper, we consider other clues. Anyway, since we only work on nouns, the POS tag is known.

3.1. Inflected forms, lemmas and stems

Whatever the method they rely on, most of WSD systems use the context of the word to be tagged through its lemmatized (Agirre & Rigau, 1995) or stemmed form (Sussna, 1993). A few others (Segond et al., 1997) use the original context (inflected forms).

In order to test which feature is the most efficient for WSD when looking at the context, the inflected form of

the word to be tagged is not taken into account for the following experiments.

3.2. Results

We have tested the KNN method on the 12 trainable nouns given in SENSEVAL. In the following table, for each noun (column 1),

- column 2 (U2) shows the number of correct assignments obtained with a unisem model (the most frequent tag is systematically assigned to the word, regardless the context).
- column 3 (F2) shows the number of correct assignments when only inflected forms are considered in training and test data.
- column 4 (L2) shows the number of correct assignments when the context is lemmatized.
- column 5 (S2) shows the number of correct assignments when all words in the context are replaced by their stems. The stemming program used is a version of the Porter stemmer (Porter, 1980).

NOUN	U2	F2	L2	S2
accident	201	237	239	240
behaviour	264	268	268	268
bet	55	142	149	146
excess	70	115	115	115
float	34	44	44	45
giant	58	79	80	82
knee	130	194	202	199
onion	181	181	181	181
promise	74	82	88	85
sack	41	49	50	51
scrap	67	90	91	91
shirt	84	151	152	154
Total	1259	1632	1659	1657
Precision	0.57	0.74	0.75	0.75

 Table 2: Comparison between inflected forms, lemmas and stems for the words in the context

This table indicates that the use of forms is less efficient than the use of lemmas and stems. The differences in global score are confirmed by the fact that for 8 (respectively 9) cases, lemmas (respectively stems) perform better than inflected forms and have the same score in the other cases.

For instance, let us consider test 700062 concerning *promise*. The word preceding *promise* in the rough test is *shows*. There is only one training sample matching this pattern (*shows promise*) and this is not enough to avoid an error: the assigned tag is 537566 (*a declaration than one will do something*). The correct tag 537626 (*the quality of potential excellence*) is assigned when using lemmas since 14 training samples are now available, corresponding to the following distribution: *show* (3), *showing* (1), *showed* (5), *shown* (4) and *shows* (1).

Globally, lemmatization slightly improves scores compared to stemming, but this is not clear when looking at each result (lemmas perform better for 3 nouns while stems perform better for 5 nouns). So, it is difficult to determine which of lemmatization and stemming performs better. These two results are coherent with known DR experiments. The variability of contexts for a particular word decreases thanks to the use of lemmas or stems and this is very important given the small amount of training data. Moreover, for DR too, the usefulness of an NLP component like lemmatization is not clear compared to the very simple stemming procedure (Lewis & Sparck-Jones, 1996).

4. Considering the inflected form of the word to be disambiguated

The inflected form of a word can be a very useful clue to find the sense. Riloff (1995) has shown that this information is important for Text Classification. Several papers describe the use of such information in their WSD systems like Ng (1997), but without measuring the exact impact of this feature. We intend to measure to what extent the inflected form is important for WSD. The following table presents the results when inflected forms, lemmas or stems are considered for the context, like in table 2, but with the inflected form and the current spelling taken into account for the word to be disambiguated. The unisem column has been suppressed. The difference between the current results and the previous ones (table 2) is indicated in order to measure the improvements.

NOUN	U3	U3- U2	F3	F3- F2	L3	L3- L2	S 3	S3- S2
accident	206	+5	242	+5	246	+7	245	+5
behaviour	264		268		268		268	
bet	104	+49	176	+34	180	+31	178	+32
excess	122	+52	146	+31	147	+32	146	+31
float	38	+4	41	-3	43	-1	43	-2
giant	90	+32	88	+11	90	+10	91	+9
knee	108	-22	188	-6	194	-6	192	-7
onion	181		183	+2	182	+1	183	+2
promise	74		86	+4	93	+5	89	+4
sack	71	+30	64	+15	65	+15	65	+14
scrap	67		100	+10	102	+11	101	+10
shirt	84		151		152		153	-1
Total	1409	+150	1733	+101	1762	+103	1754	+97
Precision	0.64	+0.07	0.79	+0.05	0.80	+0.05	0.80	+0.05

Table 3: Using inflected forms of the word to be tagged.

Firstly, this table confirms the previous results concerning the use of inflected forms, lemmas or stems for the contextual feature. Secondly, it clearly shows that the inflected form of the word to be tagged is an important clue for WSD. For example, the improvement for the word *excess* is due to the fact that there is a quasi bijection between the inflected form *excesses* and tag 512404 (*overstepping the bounds of moderation*). The lower score for *knee* indicates that the plural form of *knee* does not provide any information on its sense.

5. How to reduce ambiguity ?

It is clear that the more ambiguous a word is, the more difficult it is to disambiguate its sense. One way to reduce ambiguity is to reduce the possible number of senses. The SENSEVAL framework provides a file (mne-uid-map) in which several clues can be found concerning possible associations between senses and inflected forms or compounds.

When looking at this information, we observe that, for some words, some of their possible senses are only related to a particular use of these words. For example, sense 534544 for *scrap* is associated with the compound *scrapyard* (which can also be spelled *scrap yard* or *scrapyard*) and cannot occur in another context just as the compound *scrapyard* is always associated with sense 534544.

Therefore, the training and test data are reprocessed in order to split them into several subsets. The previous 12 tasks are now broken up into 33 subtasks. Table 4 shows the different subsets, the number of senses, samples and tests associated to each of them, and the number of good assignments for each subset when looking at the lemmatized context. Since several subsets can share a same sense, the sum of all the possible senses of each subset can be greater than the number of possible senses of the word. The last column shows the number of good assignments for unisem after splitting.

NOUN	SUB-SETS	TAGS	SAMPLES	TESTS	U4
accident		8	1265	267	201
behaviour		3	1004	279	264
bet	betting shop	1	17	48	48
	betting	7	44	92	45
	bet	7	85	134	55
excess		10	201	186	70
float	milk float	1	3	2	2
	float	11	80	73	34
giant	Giant	5	99	32	29
	giant	6	268	86	58
knee	knee breech	1	2	2	2
	knee cap	1	12	2	2
	knee hole	2	3		
	knee jerk	1	13	11	11
	knee pad	1	1		
	knee sock	1	1		
	knee up	1	1	3	3
	knee	15	427	233	129
onion	spring onion	2	2	15	8
	onion	3	39	199	181
promise		8	616	113	74
sack	sacking	3	47	32	30
	sack	8	66	50	50
scrap	scrap book	2	10	2	2
	scrap heap	2	5	6	5
	scrap metal	2	5	19	8
	scrap yard	1	7	4	4
	scrap	7	40	125	67
shirt	shirt sleeve	2	17	5	1
	sweat shirt	2	21	1	1
	T shirt	2	136	73	73
	shirt	3	377	105	83

Table 4: Split data.

When there are on average 12.5 senses for 12 nouns, there is now an average of 4.03 senses for 32 entities. If we consider the occurrences in the training corpus, there are 8.43 senses per word to be disambiguated before and 6.6 after the split. Therefore, there is a significant decrease

in ambiguity. Table 5 shows a comparison between the results obtained when splitting the sets of senses according to the use of a word and those given in table 2.

NOUN	U5	U5-	F5	F5-F2	L5	L5-	S5	S5-
		U2				L2		S2
accident	201		237		239		240	
behavior	264		268		268		268	
bet	148	+93	190	+48	193	+44	189	+43
excess	70		115		115		115	
float	36	+2	46	+2	46	+2	47	+2
giant	87	+29	91	+12	93	+13	93	+11
knee	147	+17	199	+5	207	+5	204	+5
onion	189	+8	185	+4	185	+4	185	+4
promise	74		82		88		85	
sack	71	+30	70	+21	71	+21	70	+19
scrap	86	+19	97	+6.5	98	+7	98	+7
shirt	158	+74	158	+7	159	+7	160	+6
Total	1531	+272	1738	+105.5	1762	+103	1754	+97
Precision	0.70	+0.13	0.79	+0.05	0.80	+0.05	0.80	+0.05

Table 5 : Results when splitting data

The improvement is really impressive if we consider unisem. This is logical because the splitting operation can use the previous or the following word. Thus, it is not a unisem model anymore. The results obtained with this method approximate to the ones obtained when using the inflected form of the word to be disambiguated. It is interesting to use both methods.

Let us consider the noun *bet* as an example. Out of the 44 tests that are not correctly tagged in table 2, 39 concern the inflected form *betting*. For 32 cases out of 39, a tag incompatible with *betting* was assigned by the KNN without splitting. After the split, the algorithm chose the correct sense.

6. Combining both approaches

Table 6 shows the results of a tagging obtained when splitting words into several sets and taking into account the inflected form of the word to be tagged. The comparison is made with table 5.

NOUN	U6	U6- U5	F6	F6- F5	L6	L6- L5	S6	S6- L5
accident	206	+5	242	+5	246	+5	245	+5
behaviour	264		268		268		268	
bet	148		196	+6	200	+3	197	+8
excess	122	+52	146	+31	147	+32	146	+31
float	40	+4	43	-3	45	-1	45	-2
giant	90	+3	90	-1	92	-1	92	-1
knee	125	-22	194	-5	199	-8	198	-6
onion	194	+5	190	+5	189	+4	190	-5
promise	74		86	+4	93	+5	89	+4
sack	71		65	-5	66	-5	66	-4
scrap	86		110	+13.5	112	+14	111	+13.5
shirt	158		160	+2	161	+2	162	+2
Total	1578	+47	1790	+52.3	1818	+56	1809	53.5
Precision	0.72	+0.02	0.81	+0.02	0.83	+0.03	0.82	+0.02

Table 6: Results when splitting data and using inflectedform for the word to be disambiguated.

The great improvement for *excess* is due to the fact that the form *excesses* is related to specific senses.

These results show great improvement compared to the one presented in table 2.

As a final observation, we can say that the precision obtained with lemmas (0.83) is very good compared to the ones obtained² by the different participants during SENSEVAL. It is the second best result.

7. Conclusion and perspectives

In this paper, we have shown that few easily extractable clues can help WSD. We have measured the improvement (+0.05 for the precision) obtained when taking into account the inflected form of the word to be tagged. Furthermore, we have shown that it is possible to dramatically reduce the ambiguity in many cases. The last results are really encouraging since they are obtained with a very simple method.

These methods can also be applied to other POS considered in SENSEVAL (verbs and adjectives). However, improvement is likely to be less spectacular than for nouns since adjectives have no inflected forms and the inflected verb mostly indicates tense, number or person.

The method used is very simple and there are many ways to improve the basic results. First, it is possible to assign weights according to the position of the word in the context. These weights can be calculated using a method described in (El-Bèze *et al.*, 2000). Another possibility is to expand the window of the context. Considering only 2 words may not be not sufficient for WSD. For example, Yarowsky (1992) analyses a 50-word window.

Lastly, we have performed the SENSEVAL task using a method (Loupy *et al.*, 2000) based on Semantic Classification Trees (SCT) (Kuhn & De Mori, 1995). Since SCT perform better than KNN for this task, we can expect to further improve the results by taking advantage of the clues detailed in this paper. Moreover, it should be interesting to use SCT in order to automatically detect what the useful features are (inflected forms, lemmas, etc.). This way, the knowledge found in the lexical resource provided by SENSEVAL could be learned automatically from the training corpus.

8. References

- Agirre E., Rigau G., 1995, A proposal for word sense disambiguation using conceptual distance. *First International Conference on recent advances in NLP*, Bulgaria.
- Atkins S., 1993, Tools for computer-aided corpus lexicography: the Hector project. *Acta Linguistica Hungarica* D. Beeferman, A. Berger, J. Lafferty, 1997, A model of lexical attraction and repulsion. *ACL/EACL'97*.
- El-Bèze M., Michelon P., Pernaud R., 2000, An integer programming approach to word sense disambiguation. submitted to *European Journal of Operational Research*.

² The detailed results, for each word and each participant, can be found at http://www.itri.brighton.ac.uk/events/ senseval/RESULTS/senseval.html

- Gonzalo J., Verdejo F., Chugur I., Cigarran J., 1998, Indexing with WordNet synsets can improve text retrieval. 17th International Conference On Computational Linguistics and the 36th Annual Meeting of the Association for Computational Linguistics - Workshop on Usage of WordNet for Natural Language Processing.
- Ide N., Véronis J., 1998, Introduction to the Special Issue on WSD: The State of the Art. *Special Issue on WSD*, *Computational Linguistics*, Vol.24, No. 1: p 1-40.
- Karov Y., Edelman S., 1997 ?, Similarity-based word sense disambiguation. *Special Issue on WSD*, *Computational Linguistics*, Vol.24, No. 1: p 41-60.
- Kilgarriff A., 1998, SENSEVAL: An exercice in evaluating word sense disambiguation programs. *First International Conference on Language Resources & Evaluation*: pp. 581-585, Grenade, Espagne.
- Kilgarriff A., Palmer M., 2000, Introduction to the special issue on SENSEVAL. to appear in *Computer and the Humanities, SENSEVAL special 34 (1-2)*.
- Kuhn R., De Mori R., 1995, The Application of Semantic Classification Trees to Natural Language Understanding. *IEEE, Transactions on Pattern Analysis and Machine Intelligence, vol 17, n 5.*
- Lewis D., Sparck Jones K, 1996, Natural language processing for information retrieval, *Communications of the ACM*, Vol. 39, No. 1: 92-101.
- Loupy C. de, El-Bèze M., Marteau P.-F., 1998, Word Sense Disambiguation using HMM Tagger. *First International Conference on Language Resources & Evaluation*: pp. 1255-1258, Grenade, Espagne.
- Loupy C. de, El-Bèze M., Marteau P.-F., 2000, Using semantic classification trees for word sense

disambiguation. to appear in Computer and the Humanities 34 (1-2), SENSEVAL Special.

- Ng H.T., 1997, Examplar-based word sense disambiguation : some recent improvements. *in Proceedings of the Second Conference on Empirical Methods in Natural Language Processing (EMNLP-2).*
- Porter M.F., 1980, An algorithm for suffix stripping. *Program 14 (3)*: pp. 130-137.
- Riloff E., 1995, Little words can make big difference for text classification. actes de 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval: pp. 130-136, Seattle, Washington, USA.
- Segond F., Schiller A., Grefenstette G., Chanod J.-P., 1997, An experiment in semantic tagging using hidden markov model tagging. ACL/EACL Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications, Madrid.
- Sussna M., 1993, Word sense disambiguation for free-text indexing using a massive semantic network. *in Proceedings of the second international conference on Information and knowledge management (CIKM'93)*: pp. 67-74.
- Wilks Y., Stevenson M., 1996, The grammar of sense : is word-sense tagging much more than part-of-speech tagging ?. *Report No. CS-96-05, University of Sheffield*, Sheffield.
- Yarowsky D., 1992, Word sense disambiguation using statistical models of Roget's categories trained on large corpora. *in Actes de COLING*'92: pp. 454-460, Nantes, France.